基于时序数据库的分布式网络波动监控系统

摘　要：介绍大规模数据中心应用系统间通信经常跨越多个机房或者多个核心网络区域，网络通信质量波动大幅增加，采用相应的实时监控技术架构，建立在全网网络波动监控基础平台，实现对可用性、可靠性和用户体验进行网络波动监控。

关键词：时序数据库；数据中心；信息网络；数据监控
中图分类号：TP393　　文献标识码：A
文章编号：1671-0134（2018）03-036-02　　DOI：10.19483/j.cnki.11-4653/n.2018.03.014

文 / 柴亚刚

1. 业务场景及需求

随着数据中心系统规模的不断扩大，应用系统间通信经常跨越多个机房或者多个核心网络区域，网络通信质量波动的概率大幅增加，直接影响应用的正常运行。因此，全网网络波动监控就成为基础平台监控中不可或缺的基础环节。网络波动监控主要是对可用性、可靠性和用户体验进行监测，具体为以下几方面：（1）选取每个网络区域的随机节点作为采集点，同时对跨机房和跨网络区域的 ICMP 探测响应往返时间进行持续监测。（2）选取用户端网络节点作为采集点，对所有服务域名和关键 URL 的响应时间、响应状态和响应内容等持续监测。（3）能够对历史数据进行实时查询，能够对一定时期内的监测指标进行聚合计算，通过各类图表形式展示网络响应。（4）能够定义告警策略，当采集指标符合告警策略时，通过 Web Hook、Email 等方式进行告警。网络波动监控的数据类型主要是时间序列数据，因此，考虑用时序数据库，配合分布式采集工具、消息系统以及监控前端系统来实现。

2. 时序数据库的当前发展

与普通数据不同，每条时间序列数据都带有一个时间戳，反映的是某个时间点的度量情况。时间序列数据库（Time Series Database）则是针对时间戳或时间序列数据而优化的数据库，专门用于跟踪、监控、聚合和处理随时间变化的度量或者事件，这里的度量可以是服务器指标、应用性能监控数据、网络性能数据、传感器数据、事件、点击、市场交易以及其他各类数据。

时间序列数据库并不是新生事物，但其早期主要用于交易系统，用来监测股票交易的波动性。然而，过去十年中，随着 PC 服务器逐步替代大型机和小型机，互联网、物联网和大数据技术飞速发展，各类时间序列数据、指标和事件随时随地、无时无刻在产生，随着数据源的变化而衍生的对数据生命周期管理、长时间跨度下实时快速查询和聚合计算、根据历史数据对未来趋势进行预判等新的业务需求，要求底层数据基础架构也随之变化，需要更适合互联网的分布式时间序列数据库。

在时间序列数据库领域，InfluxDB、RRDtool、Graphite 和 OpenTSDB、Druid、Prometheum 的排名比较靠前，使用也更为广泛。如下图所示。

22 systems in ranking, November 2017

Rank Nov 2017	Rank Oct 2017	Rank Nov 2016	DBMS	Database Model	Score Nov 2017	Score Oct 2017	Score Nov 2016
1.	1.	1.	InfluxDB	Time Series DBMS	9.34	+0.63	+3.74
2.	2.	2.	RRDtool	Time Series DBMS	3.19	+0.08	+0.72
3.	3.	3.	Graphite	Time Series DBMS	2.86	+0.09	+0.95
4.	↑5.	↑5.	Kdb+	Multi-model	1.85	+0.02	+0.68
5.	↓4.	↓4.	OpenTSDB	Time Series DBMS	1.71	-0.15	+0.25
6.	6.	6.	Druid	Time Series DBMS	0.98	-0.02	+0.35
7.	7.	7.	Prometheus	Time Series DBMS	0.81	+0.07	+0.50
8.	8.	8.	KairosDB	Time Series DBMS	0.46	-0.02	+0.19
9.	9.	9.	eXtremeDB	Multi-model	0.30	-0.02	+0.12
10.	10.	10.	Riak TS	Time Series DBMS	0.21	-0.03	+0.08
11.	↑12.	↑19.	Hawkular Metrics	Time Series DBMS	0.16	+0.02	+0.16
12.	↑13.	↑16.	Blueflood	Time Series DBMS	0.16	+0.02	+0.14
13.	↓11.	↑15.	Axibase	Time Series DBMS	0.10	-0.09	+0.08
14.	14.	↑18.	Machbase	Time Series DBMS	0.06	-0.01	+0.06
15.	↑16.	↑17.	TempoIQ	Time Series DBMS	0.06	+0.02	+0.05
16.	↓15.	↓13.	Warp 10	Time Series DBMS	0.05	-0.01	+0.01
17.	17.	↓11.	Heroic	Time Series DBMS	0.02	+0.02	-0.07
18.			IRONdb	Time Series DBMS	0.00		
18.	↓17.	↓14.	Newts	Time Series DBMS	0.00	±0.00	-0.03
18.	↓17.		SiriDB	Time Series DBMS	0.00	±0.00	
18.	↓17.	↑19.	SiteWhere	Time Series DBMS	0.00	±0.00	±0.00
18.	↓17.	↓12.	Yanza	Time Series DBMS	0.00	±0.00	-0.06

时序数据库 11 月份排名情况（摘自 db-engines.com[1]）

3. 时序数据库的几个关键概念

以使用最广泛的 Influxdb 为例，有以下几个关键概念：

（1）field key/field value/field set，度量指标数据，field key 为度量指标字段，filed value 为对应的值，两者构成 field set。Field key 没有索引，基于 field 的过滤查询都是全表扫描。Field value 的值类型只能为字符串、浮点数、整型数或者布尔类型。每条度量指标数据都和一个时间戳绑定。

（2）tag key/tag value/tag set，可选的索引标签，tag key 为索引标签字段，tag value 为对应的值，两者构成 tag set。在查询语句中，可以跟在 where 短语后面。

（3）measurement，类似于关系型数据库中的表，存放 tags、fields 以及对应的时间戳。

（4）retention policies，数据存储策略，默认为永久保存，可以为数据表设置过期时间，influxdb 会定期清理。

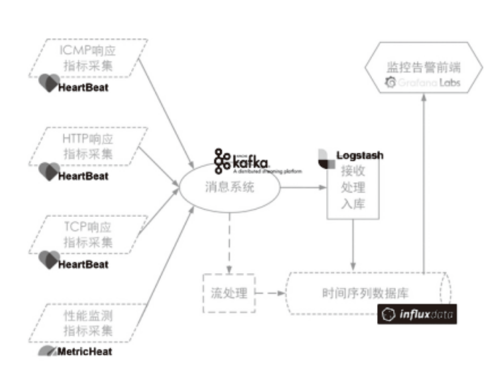
（5）database，类似于关系型数据库中的数据库，

逻辑概念，包含用户权限、存储策略、时序数据等。

（6）series，数据序列，相同的数据表、存储策略和 tag set，构成 1 个 series。一般情况下，表中有 tag 标签时，根据不同 tag 标签的排列组合会形成多条数据序列。这是时序数据库中最关键的概念。

4. 实时监控技术架构

由于采集网络范围广、节点多、采集时间间隔短、数据插入并发高，因此，网络波动指标数据的采集处理选用分布式架构。在全网采集点上部署采集工具，分别以 JSON 标准格式上传到消息系统，接收处理入库程序，从消息系统接收消息，处理后插入时间序列数据库。前端监控系统从时间序列数据库读取数据，并在前台进行展示，依据定制策略进行告警。如下图所示。



网络波动监控系统架构

采集工具选用 HeartBeat，属于 Elastic Stack 中的轻量型数据采集器 Beats 工具集中的一种。它采用 Go 语言开发，并发性能较高，支持 icmp/http/tcp 三种类型的心跳监控，动态地添加和删除目标，既支持直接输出到 ElasticSearch 和 Logstash，也支持输出到 Kafka 和 Redis 消息队列。消息系统选用 Kafka，属于 Apache 基金会项目，被定义为分布式流处理平台，通常用于实时流数据的管道或者流数据处理应用。它采用 Java 语言开发，并发性能高，支持发布 / 订阅的消息系统场景以及日志存储场景。

时序数据库选用 InfluxDB，是 InfluxData 所开源的项目。在 db-engines.com 所公布的时序数据库中长期排名第一，是目前应用最广泛的时序数据库。它支持对时序数据定期存储，可以根据数据量和时间定时清理过期数据，避免磁盘空间超标；支持对时序数据进行 mean/min/max/last/first/avg 等各种快速聚合计算。

监控告警前端选用 Grafana，支持 InfluxDB、ElasticSearch、Graphite、Prometheus 等各类数据源，支持图表、表格、仪表盘等各类展示方式，通过自定义告警水位和告警信息实现告警。

5. 时序数据存储与查询

网络波动以 ICMP 和 HTTP 响应结果指标数据为主，分别存储在不同数序数据库中，每个数据库中根据数据来源、时间等字段适当拆表，同时对响应状态错误的结果数据复制一份入库到错误表，方便查询。

由于是对网络波动度量指标数据进行监测，因此，field 数据主要是 ICMP 和 HTTP 请求的响应时长数据，将采集器名称、监测目标、返回状态等作为 tag 数据。

ICMP 响应的指标数据仅有一个 rtt 时长，HTTP 响应的指标数据较为复杂。各列数据的含义为：time 为时间戳，wreponse_status 表示响应状态，wup 表示是否有响应，wresolve_rtt 表示 DNS 解析时间，wtcp_connect 表示 tcp 连接时间，whttp_rtt 表示 http 响应时。

Influxdb 的查询语法与 SQL 类似，增加和优化了对一定时间范围内的指标数据进行最大值 / 最小值 / 平均数计算。以网络区域间 ICMP 响应序列查询为例，选取某个区域节点的 ICMP 响应时长的平均值进行聚合查询，用时间间隔作为聚合依据。以下为查询语句。

```
SELECT max("wduration") FROM "RF-platfapp-ping-HeartBeat" WHERE "wmonitor" =~
/^$montarget$/ AND "wrf_observer" =~ /^$monobserver$/ AND $timeFilter GROUP BY
time($moninterval) fill(previous)
```

相较于网络区域间 ICMP 响应监控，重点网站的 URL 监控的 HTTP 响应数据的指标更多，包括 TCP 连接时间、DNS 解析时间、HTTP 响应时间，HTTP 响应状态。以下为查询语句：

```
SELECT mean("wduration") AS "响应时间", mean("wresolve_rtt") AS "DNS 解析时间",
mean("wtcp_connect_rtt") AS "TCP 连接时间", mean("whttp_rtt") AS "HTTP 响应时间"
FROM "autogen"."RF-WWW-HeartBeat" WHERE "wmonitor" =~ /^$montarget$/ AND
$timeFilter GROUP BY time(10m) fill(previous)
```

6. 总结与展望

本文从跨机房和网络区域的网络质量波动监控实际需求出发，设计了基于时序数据库的分布式网络波动监测系统。通过消息队列系统实现指标数据的管道传输，使得分布部署在不同机房和网络区域的采集节点和数据接收处理模块解耦，有效地扩大了监控范围和容量；通过时序数据库来存储网络质量指标数据，通过前端展示组件实现指标数据的图表展示，有效解决了网络质量和波动的实时监测和历史数据查询的可视化监控需求。媒

参考文献

[1] 林芝．基于信息论网络的时序数据库挖掘 [J]．计算机工程与应用，2003（01）．

[2] 黄河．时序数据库中快速相似搜索的算法研究 [J]．模式识别与人工智能，2003（02）．

[3] 郭四稳．基于小波技术的网络时序数据挖掘 [J]．计算机工程，2007（02）．

（作者单位：新华社辽宁分社）